機械学習による薬物分子-ターゲット相互作用予測

AUTHOR(S):

馬見塚, 拓

CITATION:

馬見塚, 拓. 機械学習による薬物分子-ターゲット相互作用予測. SAR News 2015, 29: 2-8

ISSUE DATE:

2015-10

URL:

http://hdl.handle.net/2433/218531

///// **Perspective/Retrospective** /////

機械学習による薬物分子－ターゲット相互作用予測

京都大学化学研究所　馬見塚　拓

1. はじめに

薬物分子（以後、簡便のため薬）や薬の候補である低分子化合物と、それらのターゲットであるタンパク質の相互作用を予測することは、薬科学の中心課題である創薬はもちろん、薬の副作用やリポジショニング等、薬に関連する様々な問題の解決に寄与可能な課題である。相互作用は、実際の実験により測定可能で、また実験が最も信頼性が高いことはもちろんである。しかし現在、薬の候補となる低分子化合物は 3,000 万種類以上が PubChem 等の低分子化合物のデータベースに格納されており、一方、ターゲットであるタンパク質の候補は、ヒトの遺伝子で 1 万以上に上り、非常に多い。従って、これらの組み合わせを網羅的に実験することは、時間及びコストの観点から非常に難しい。そこで、この問題を克服するために、現在まで、計算機を使った相互作用の予測が様々な角度から試みられてきた [1]。計算機による手法は、大まかに : 1) シミュレーションと 2) 機械学習（データマイニング）に分けられる。

シミュレーションは、電子状態等物理化学上の性質を深く考慮したものから、立体障害のみを考慮するものまで様々なレベルがある。いずれのレベルにおいてもシミュレーションを行うためには、薬（低分子化合物）とターゲット、両者の立体（3 次元）構造が必要になる。特に、立体構造既知のターゲット、すなわちタンパク質立体構造データベース PDB（Protein Data Bank）にあるタンパク質は限られており、さらに解かれているタンパク質の分布には偏りがある。実際、代表的なターゲットである GPCR（G-Protein Coupled Receptor）の立体構造は、近年増加傾向にあるとはいえ、これまで数十例しか報告されていない。さらに、シミュレーションは実験より効率的とはいえ、上記のような膨大な組み合わせを考慮した場合に計算時間の効率性が十分とは言い難い。

一方、近年、実験手法のハイスループット化により、薬－ターゲット相互作用データが蓄積されつつある（例えば、[2]）。これら既知の相互作用データを低分子化合物とタンパク質の相互作用予測に役立てたいと考えるのは極めて自然である。機械学習（データマイニング）は、所与のデータからデータに内在する規則や仮説を推定する手法全般を指す。非常に単純な方法から最適化問題として計算時間を要する手法まで多様であるが、一般に計算時間等のコストはシミュレーションに較べるとはるかに低い。また薬－ターゲット相互作用予測において、タンパク質の立体構造が既知である必要はない。機械学習の研究開発は 1980 年代に開始され、その当時から信号処理、音声認識、自然言語処理、生命科学等幅広い様々な分野で応用され、多くの成功を収めてきた。例えば、現在、携帯電話やタブレット等には通常音声認識アプリが組み込まれているが、元々のモデルパラメータは機械学習で推定されている。また、近年では、材料科学や機械工学等の分野でシミュレーションを高精度で近似する代替手段として機械学習を中心としたデータ科学、Materials Informatics と呼ばれる分野が生まれている等、多方面で機械学習の応用が脚光を浴びている。（機械学習の教科書として[3]。和訳もある）

さて、機械学習の観点から、薬とターゲットが相互作用するか否かを推定する問題は、2 値ラベル（クラス）の分類問題とみなせる。これは機械学習の標準的な問題で、典型的なデータ（後述する特徴ベクトル）に対しては多様な手法が既に提案されている。ただし、化合物とタンパク質は、いずれも特徴ベクトルのみならず非常に多様な情報を有している。まず化合物については、物理化学的性質を中心として、連続値や離散値のベクトルである記述子を各化合物に計算することが可能で、化合物からそのような記述子を計算する手法やツールが頻繁に利用され、特徴ベクトルとして利用できる。別の観点では、化合物がそもそも持つユニークな化学構造を用い、原子をノードとし、原子間の結合をエッジとしたグラフ（分子グラフ）で表現することもできる。また、薬に関しては、薬効分類や骨格分類等の分類がよくなされており、分類木を利用することが

可能である。例えば、分類木上の距離を使って薬同士の相同性を計算することもできる。一方、タンパク質は、そもそも遺伝子の発現であるため、遺伝子の情報も同様に利用することができる。例えば、遺伝子配列のモチーフの有無や遺伝子発現情報は、それぞれ離散値と実数値として、特徴ベクトルにできる。しかし、遺伝子配列そのものは長さが不定であり、また遺伝子機能は階層性を持つことからベクトル表現は難しい。さらに、タンパク質自身の情報として、立体構造やタンパク質間相互作用等がある。これらはいずれも特徴ベクトルとして情報を表現し切れない。従って、2 値ラベルの分類問題に対して、このような情報を有効に利用するためには、機械学習において新しい問題設定と解決手法を構築する必要があり、非常に刺激的な課題である。

本稿では、まず、薬とターゲット間の相互作用を予測する機械学習手法全般について大まかに俯瞰する。次に、それらの中でも、薬とターゲットそれぞれに対し、化合物間、タンパク質間の相同性がデータとして与えられる場合に着目する。すなわち、相互作用データのみならず、化合物間相同性およびタンパク質間相同性が、それぞれ複数与えられた場合の機械学習による解決方法を紹介し、それらの特徴を俯瞰する。この問題設定は、機械学習の問題設定としては非常にユニークだが、機械学習の様々な応用を俯瞰すれば一般性があり、薬－ターゲット相互作用予測はこの問題設定の典型的な応用例の一つである。

2. 薬－ターゲット相互作用予測のための機械学習

本章では、まず、機械学習により薬－ターゲット相互作用を予測するアプローチを大まかに 3 つに分けて取り上げる。

2.1 特徴ベクトルに基づくアプローチ

機械学習では、データの各単位を事例と呼ぶ。一般に、事例を説明する変数は属性あるいは特徴と呼ばれる。最も典型的な機械学習のデータは、各事例が固定サイズのベクトル（特徴ベクトル）の集合となる。言い換えると、このデータは表（行列）であり、各行が事例の特徴ベクトルとなり、各列は属性（特徴）である。2 値の分類問題では、属性の一つが 2 値ラベルとなり、この属性のラベルを分類できるよう、他の属性を使って学習を行う。すなわち、事例のクラスを特定の属性で表現する。いったん表が出来上がれば、様々な既存の分類手法がこの表に適用可能である。薬－ターゲット間相互作用予測問題も例外ではない。化合物とタンパク質のペアを事例とし、化合物とタンパク質それぞれから属性を作成し表ができる。例えば、属性として、化合物に対しては分子記述子、タンパク質に対しては遺伝子配列のモチーフや遺伝子発現が利用できる。ラベル属性を相互作用の有無として、このような表をいったん作れば、サポートベクトルマシン、決定木といった機械学習の分類手法を表に対して適用して、予測モデルを作ることができる [4]。（各分類手法の内容については本稿では詳述しないので、分類手法とは、特徴ベクトルと 2 値ラベル（クラス）からなる表（行列）から学習を行い、予測では特徴ベクトルを入力とし 2 値ラベルを出力できる手法と考えていただきたい）

ただし、これは、既存の機械学習の枠組みにデータを当てはめる直接的なアプローチであり、少なくとも以下の 2 つの問題がある。

1）使用する分類手法は一般的なものであり、薬―ターゲット相互作用に固有の性質を十分に反映した手法には必ずしもなっていない。すなわち、分類手法の基本的なアイデアはいずれもデータ空間上で 2 つのクラスに属するデータを判別する関数を見つけることである。一方、現実の相互作用は薬とターゲットの立体構造上起こる現象である。つまり、データ空間が立体構造上の相互作用が起こるかどうかを反映していなければ、判別関数に意味がない。

2）記述子等の属性はクラスラベルに関係なく生成されるので、分類に不要な属性が多く含まれる可能性がある。例えば、2 値を取るある記述子を使用した場合に、相互作用のある事例はほとんど一方の値しか取らないかもしれない。また、連続値の場合も相互作用がある事例とそうでない事例とでほとんど値が変わらない可能性もある。

以上の問題は、薬－ターゲット相互作用予測に特有の性質を考慮せずにデータを構築し既存分類手法を適用するが故に起こる。言い換えると化合物やタンパク質の豊富な情報を有効活用していないことに由来する。解決手法として、例えば、次の 2 点が挙げられる。

1）サポートベクトルマシンで重要なカーネル関数はデータ空間上の事例の距離を表すので、カーネル関数を設計する際に、特徴ベクトルのみでは表現し切れない、化合物やタンパク質の豊富な情報を利用する。例えば、3.1 節の Pairwise kernel method [5]では、化合物間の相同性とタンパク質間の相同性を入力として、それらを利用したカーネル関数を設計することにより、実際の立体構造上の類似性をより反映したデータ空間を構成し、その上での判別を行う。

2）化合物側の化学構造とタンパク質側の遺伝子配列に対し、実際にデータに頻繁に出現する部分構造（頻出部分構造）のみを属性として利用することも、構造に直接影響を与えかつデータに出現する属性のみを利用する有効な方法である [6]。(ここで、部分構造とは、例えば、化学構造を分子グラフとみなせば部分グラフを指し、また、遺伝子配列を文字列とみなせば部分文字列を指す)

このように、特徴ベクトルによる方法は、機械学習を適用する手っ取り早い方法ではあるが、本来のデータに内在する特徴を十分に把握できるとは言い切れない。

2.2 相同性行列を用いるアプローチ

化合物、タンパク質いずれに対しても、それらの属性ではなく、化合物間、タンパク質間の相同性を使用する。従って、3 つの行列が入力となる。すなわち、まず、行と列が化合物とタンパク質からなり、各要素が相互作用の有無を指す行列。次に、化合物間相同性行列。この場合は、行も列も化合物で、要素は相同性を示す値となる。同様に、タンパク質間相同性行列で、行と列両方ともタンパク質で要素は相同性値。ただし、相同性は、タンパク質であれば、遺伝子配列、モチーフの有無、遺伝子発現、タンパク質相互作用等、複数の相同性が得られる。従って、入力は、3 種類の行列であり、化合物相同性とタンパク質相同性に関しては、それぞれ複数の行列が入力となる [7]。

相同性行列を用いる手法が近年脚光を浴びている背景には、ケモゲノミクスやケミカルバイオロジーに由来する以下のような考え方がある：相同性（その逆は距離）は、化合物とタンパク質それぞれの空間でのデータ分布を表している。すなわち、化合物はそのデータ空間であるケミカルスペース内に分布し、それは化合物の相同性によって表現される。同様にバイオロジカルスペースはタンパク質の分布を表す空間であり、この空間内のタンパク質の分布は相同性によって表現される。その 2 つのスペースのインタフェースに薬－ターゲット間相互作用があると考えられる。従って、2 つのスペースとそのインタフェース（既存相互作用）がうまく把握されれば、薬－ターゲット相互作用の有無は、インタフェースも含めたスペース全体を使ってクリアに説明されるはずである。言い換えると、ケミカルワールドとバイオロジカルワールドが交わる全体像が見える、と考えられる [8]。なお、特徴ベクトルによるアプローチの中でも化合物側の特徴とタンパク質側の特徴を明確に区別するモデルを使用する方法があり、そのような方法は、同じ考えに基づくと言える [9]。

本稿では、3 章において、相同性行列を用いるアプローチの既存手法をより具体的に紹介するとともに、問題点を指摘し、それらを克服する手法を紹介する。

2.3 その他のアプローチ

その他のアプローチとして、例えば、生命科学文献データを入力とするテキストマイニングがある [10]。文献データベースの中で同一の文書（あるいは章、パラグラフ等）に、対象とする化合物とタンパク質が共に出現するという共起情報から、薬－ターゲット相互作用を予測する。このアプローチの問題は、共起が実際の相互作用を反映するとは限らないことである。むしろ探索的に未知の薬－ターゲット相互作用を検出／予測するアプローチである。

3. 相同性行列を用いるアプローチ

本節では、2.2 で述べた相同性行列を用いる手法を俯瞰する。繰り返しになるが、このアプローチでは、入力は 3 種類の行列である:薬―ターゲット相互作用を示す行列 Y、化合物間相同性を示す行列 S_d、タンパク質間相同性を示す行列 S_t（S_d と S_t は入力行列であることに注意）。化合

物とタンパク質それぞれの空間を適切に利用し、相互作用を予測できるモデルを構築することを目指す。

3.1 Pairwise kernel method (PKM)

2.1 節で説明したように、事例を化合物とタンパク質のペアとし、サポートベクトルマシン等の分類手法を適用する [5]。サポートベクトルマシンではカーネル関数を必要とするが、既存の特徴ベクトルから相互作用予測に適切なカーネル関数が計算できるかどうかは不明である。そこで、入力である相同性行列を使いカーネルを計算する。すなわち、化合物とタンパク質のあるペア P_A とあるペア P_B の相同性は、P_A の化合物と P_B の化合物の相同性、P_A のタンパク質と P_B のタンパク質の相同性を単純に使うことによって得られる。考え方は合理的で簡単に計算できるが、この方法ではすべてのペアの組み合わせに対してカーネル関数を計算する必要がある。つまり、例えば、10,000 個の化合物と 1,000 個のタンパク質に対して 10,000×1,000 の化合物とタンパク質のペアがあり、これらの組み合わせなので、10,000×1,000×10,000×1,000 というメモリ空間が必要になり、メモリ空間のスケーラビリティに問題がある。

3.2 Bipartite local models (BLM) と変形手法

PKM は、ペアの組み合わせを考えることによりスケーラビリティの破たんをきたした。そのためペアを避け、化合物側とタンパク質側をそれぞれ独立して考える（BLM:Bipartite local models）[11]。具体的には、ある化合物 C とあるタンパク質 P の相互作用を予測する際に、まず化合物側を C に固定し相互作用をサポートベクトルマシンにより予測する。すなわち、各タンパク質を事例として、化合物 C とタンパク質の相互作用の有無をラベルとし、タンパク質間の相同性をカーネル関数としてサポートベクトルマシンにより、化合物 C とタンパク質 P の相互作用を学習・予測する。次に同様にタンパク質側を P に固定し、今度は各化合物を事例として、化合物間の相同性をカーネル関数として相互作用をサポートベクトルマシンの学習により予測する。最終的な予測は、2 つのサポートベクトルマシンの予測を組み合わせる。この手法は、特徴ベクトルの代わりに相同性をうまく利用した手法で、PKM のメモリ空間のスケーラビリティの問題を回避できる。しかし、2 つの問題がある：1）各相互作用を予測する毎にサポートベクトルマシンを 2 回学習する必要があり、例えば、10,000 個の化合物と 1,000 個のタンパク質に対して 2×10,000×1,000 回ものサポートベクトルマシンの学習と予測が必要となり計算時間（計算量）がかかる。2）クラスラベルに相互作用情報が利用されるのを除けば、データが化合物側とタンパク質側で全く独立に扱われる。

BLM の変形手法は数多く提案されており、例えば、サポートベクトルマシンの代わりに線形回帰を行い、正則化項に相同性行列を利用する手法（NetLapRLS [12]）や、PKM の考え方を少しでも取り入れようとする手法（GIP [13]）等がある。しかしいずれにせよ、これらの手法のフレームは BLM と同じであり、そのため上記 2 つの問題点が残る。

3.3 行列分解による手法

上記の 2 つのアプローチのそれぞれの問題点を踏まえて、計算量やメモリ空間は抑えつつも、化合物とタンパク質のデータを（独立ではなく）融合的に扱い予測モデルを構築するという手法である [14,15]。具体的には、化合物×タンパク質の相互作用行列 Y を低次元行列 A、B に分解する。つまり、A、B はそれぞれ化合物側、タンパク質側の低次元行列であり、この時、A、B がそれぞれ化合物間、タンパク質間の相同性行列 S_d、S_t を再構築できるように分解する。言い換えて、逆の方向から説明をすれば、相同性行列 S_d、S_t を分解することによって得られる低次元行列 A、B（つまり $S_d=AA^T$、$S_t=BB^T$：T は行列の転置を意味する）が相互作用行列 Y を構成するように（$Y=AB^T$）A と B を求めるのである。つまり、化合物およびタンパク質の相同性行列を分解することによってそれぞれ得られる A、B は、言わばケミカルスペースとバイオロジカルスペースのエッセンス（因子）であり、これらエッセンスで相互作用が説明できるようにパラメータ A、B を推定する。エッセンスである低次元行列 A、B から、ケミカルスペース S_d とバイオロジカルスペース S_t、さらに相互作用行列 Y という 3 つの入力行列すべてを構築（説明）できるように A、B を推定する。この方法は、2 種類の相同性行列と相互作用行列を融合的に使用

し相互作用を説明する。考え方が合理的で、大きな計算量やメモリ空間を必要としない。しかも、相互作用を説明する際の、ケミカルスペースとバイオロジカルスペースのエッセンスをも表示可能という利点を持つ。

3.4 性能比較

薬ーターゲット相互作用予測に対し相同性行列を用いる手法の性能比較を [15]から抜粋して表 1 に示す。評価値は precision-recall 曲線（実数値で与えられる予測により事例をソートし、ある閾値を予測実数値に対して置いた時に、閾値以上の事例の中で正例の予測率（precision）とカバー率（recall）を、閾値をずらしながら Y,X 軸にプロットしたもの）の下の面積を指す AUPR（Area Under the Precision Recall curve）であり、機械学習の分類性能評価で最も使われる AUC（Area Under the ROC Curve）に較べて、予測上位に対する性能評価により重きがおかれる。データはベンチマークとして最も使われる 4 種類のデータセットを用いている。評価手順は交差検証（クロスバリデーション）である。上記 3.3 章で紹介した行列分解による手法が CMF と MSCMF であり（CMF は化合物側、タンパク質側、それぞれ一つの相同性行列のみ用いた場合の結果）、この表から相同性行列を用いる手法の性能面での優位性が窺える。

表 1. 相同性行列を用いる手法の AUPR による性能比較。

方法		NUCLEAR RECEPTOR	GPCR	ION CHANNEL	ENZYME
PKM	**(3.1 節)**	0.514	0.474	0.663	0.627
BLM	**(3.2 節)**	0.204	0.464	0.592	0.496
NETLAPRLS	**(3.2 節)**	0.563	0.708	0.9	0.874
GIP	**(3.2 節)**	0.604	0.727	0.898	0.884
CMF	**(3.3 節)**	0.643	0.746	0.937	0.887
MSCMF	**(3.3 節)**	**0.673**	**0.773**	**0.937**	**0.894**

また、MSCMF では、複数の相同性行列に対する重みを学習可能で、重みから相同性行列の予測への寄与を測ることができる。表 2 に結果を示す [15]。GO は Gene Ontology 上の距離から計算した相同性であり、MF は Gene Ontology の Molecular Function のカテゴリの情報を用いた距離, 同様に BP は Biological process のカテゴリ情報により相同性を計算している。また、PPI はタンパク質間相互作用ネットワーク上の距離から計算した相同性である。この結果から配列相同性の寄与は GPCR を除き意外に低いことや、GO の 2 つの相同性はほぼ同様に予測に寄与することが窺える。

表 2. ターゲットの相同性行列に対する重みの学習結果の典型例

相同性行列	NUCLEAR RECEPTOR	GPCR	ION CHANNEL	ENZYME
配列相同性	0	0.5297	0	0
GO (MF)	0.4409	0.1286	0.5262	0.3827
GO (BP)	0.5591	0	0.4738	0.3652
PPI	0	0.3417	0	0.2521

4. おわりに

薬とターゲットの相互作用予測に対する機械学習手法を俯瞰し、特に化合物とタンパク質の相同性行列を用いるアプローチに関して、代表的な手法とその比較を行った。相同性を有効に使うことにより、予測精度の向上のみならず、予測に重要な要因と予測に有効な相同性を知ることが可能になる。このアプローチでは、相同性を空間と同様に扱い、薬とターゲットの空間を相互作

用という観点から統合的に扱う。一方で、現在既知の相互作用は非常に疎であり、相同性行列と相互作用の空間を構成するよう推定した低次元行列（因子）が、既存データにオーバーフィットしやすいという問題（所謂 me-too-drug の問題）があり得る。これはデータ獲得コストが高くデータ量が十分でない場合にしばしば見られる問題だが、今後、実験面でのハイスループット技術の進展等によるデータや知見の増加と、同時に、そのような問題をより考慮した機械学習技術の革新による解決が望まれる。

最後に、これは読者の興味からやや外れるかもしれないが、機械学習にとって、今回紹介した相同性の問題設定は一般性があり、様々な応用に適用可能であることを述べたい。特に、データマイニングにおける典型的な問題、レコメンデーションも対象となる。レコメンデーションでは、ユーザ集合と商品集合があり、ユーザと商品間の購買情報がある。既知購買情報を利用して、未知の購買情報を推定する。これは、化合物集合とタンパク質集合があり、化合物とタンパク質間の既知相互作用から未知相互作用を予測する問題と同じである。さらに、化合物間相同性、タンパク質間相同性と同様に、ユーザ間、商品間の類似性情報から相同性を入力可能である（特徴ベクトルも可能）。従って、薬―ターゲット相互作用予測問題は、レコメンデーションの問題と同じである。今後、薬―ターゲット相互作用予測への機械学習技術の革新がレコメンデーション技術の進歩につながる可能性があり、また、逆も真である。実際に MSCMF と同様の行列分解の技術がレコメンデーション（協調フィルタリング）でも提案・利用されている [16]。ただ、実データの観点からみると違いがある。すなわち、レコメンデーションでは、ユーザ数は数百万～数千万、商品も数十万に及ぶことさえあるビッグデータである。一方、タンパク質は 1 万程度の種類に留まる。従って、データサイズに違いがあり、実際、薬－ターゲット相互作用予測に非常に有効な行列分解は、レコメンデーションに対しては、よりスケーラビリティの改善が要求される。いずれにせよ、このような複数分野での応用の存在が様々な刺激となり、技術革新がより進むと予想される。

謝辞

紹介した研究の中で、著者が含まれる研究は、多くの共同研究者の方々による成果である。特に、中国復旦大学の Shanfeng Zhu 先生、彼の研究室の学生であった Xiaodong Zheng さん、Hao Ding さん、また、北海道大学の瀧川一学先生、東京大学の津田宏冶先生には、多岐に渡りご教示いただいた。彼らの努力により、私を含めて関連分野の多くの方々が新しい知見を得ることが出来た。研究推進には、JST BIRD、科研費＃24300054、京都大学化学研究所若手研究者受入事業、京都大学化学研究所共同利用・共同研究課題＃2014-27、2015-33 のサポートを受けた。最後に、執筆の機会を与えていただいた SAR News 関係者の方々に厚く御礼申し上げる。

参考文献

[1] Hopkins, A. Drug discovery: predicting promiscuity, *Nature*, **462**(7270), 167–168 (2009).
[2] https://www.ebi.ac.uk/chembl/
[3] Hastie, T., Tibshirani, R. and Friedman, J. The Elements of Statistical Learning, Springer (2009).
[4] Nagamine, N. and Sakakibara, Y. Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data, *Bioinformatics*, **23**(15), 2004–2012 (2007).
[5] Jacob, L. and Vert, J-P. Protein-ligand interaction prediction: an improved chemogenomics approach, *Bioinformatics*, **24**(19), 2149–2156 (2008).
[6] Takigawa, I., Tsuda, K. and Mamitsuka, H., Mining significant substructure pairs for interpreting polypharmacology in drug-target network, *PLoS One*, **6**(2), e16999 (2011).
[7] Ding, H., Takigawa, I., Mamitsuka, H. and Zhu, S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review, *Briefings in Bioinformatics*, **15** (5), 737-747 (2014).
[8] Lipinski, C. and Hopkins, A. Navigating chemical space for biology and medicine, *Nature*, **432**(7019), 855-861 (2004).
[9] Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G., and Okuno, Y. Analysis of multiple compound–protein interactions reveals novel bioactive molecules, *Molecular Systems Biology*, **7**, 472. doi:10.1038/msb.2011.5 (2011).

[10] Zhu, S., Okuno, Y., Tsujimoto, G. and Mamitsuka, H. A probabilistic model for mining implicit "chemical compound - gene" relations from literature, *Bioinformatics*, **21**(Suppl 2), ii245-ii251 (2005).

[11] Bleakley, K. and Yamanishi, Y. Supervised prediction of drug-target interactions using bipartite local models, *Bioinformatics*, **25**(18), 2397–2403 (2009).

[12] Xia, Z., Wu, L-Y., Zhou, X., and Wong, S. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces, *BMC Systems Biology*, **4**(Suppl 2), S6 (2010).

[13] van Laarhoven, T., Nabuurs, S. B. and Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction, *Bioinformatics*, **27**(21), 3036–3043 (2011).

[14] Gönen, M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization, *Bioinformatics*, **28**(18), 2304–2310 (2012).

[15] Zheng, X., Ding, H., Mamitsuka, H. and Zhu, S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *ACM SIGKDD*, 1025–1033. ACM (2013).

[16] Gu, Q., Zhou, J. and Ding, C. H. Q. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In *SDM*, 199–210. SIAM (2010).